Corpus-based learning of Cantonese for Mandarin speakers

John Lee[1] and Tak-Sum Wong[2]

Abstract. This paper reports our experience in using a parallel corpus to teach Cantonese, a variety of Chinese spoken in Hong Kong, as a second language. The parallel corpus consists of pairs of word-aligned sentences in Cantonese and Mandarin Chinese, drawn from television programs in Hong Kong (Lee, 2011). We evaluated our pedagogical approach with Mandarin-speaking students at a university course. For each student, we first diagnosed the set of Cantonese words with which s/he experienced difficulties. Then, on a web-based interface, the student independently searched in the parallel corpus for sentence pairs involving this set of Cantonese words, and analysed the translations and usage examples. Our experiments showed that, in both the short- and long-term, the corpus-based pedagogical method helped students better retain their knowledge of difficult Cantonese words.

Keywords: parallel corpus, language acquisition, Cantonese, Mandarin.

1. Introduction

Since its return to China in 1997, Hong Kong has received a large number of visitors from mainland China to study and work in the city. There has thus been a marked increase in the need to teach Cantonese, the Sinitic variety spoken in Hong Kong, to the mainland Chinese, most of whom speak Mandarin Chinese as their first language. Since both languages are developed from Middle Chinese, they share many cognates with strong, regular phonological correspondence. Nonetheless, they are not mutually intelligible.

1. Department of Linguistics and Translation, City University of Hong Kong; jsylee@cityu.edu.hk.

2. Department of Linguistics and Translation, City University of Hong Kong; tswong-c@my.cityu.edu.hk.

How to cite this article: Lee, J., & Wong, T.-S. (2014). Corpus-based learning of Cantonese for Mandarin Speakers. In S. Jager, L. Bradley, E. J. Meima, & S. Thouësny (Eds), *CALL Design: Principles and Practice*; *Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands* (pp. 196-201). Dublin: Research-publishing.net. doi:10.14705/rpnet.2014.000217

Spoken by more than 55 million people, Cantonese is the "most widely known and influential variety of Chinese other than Mandarin" (Matthews & Yip, 2011, p. 3). Because Cantonese is a predominantly spoken language, it is relatively difficult for learners to find written samples of the language. Example sentences in textbooks are often artificially created, and do not always reflect the most colloquial or current usage. In this paper, we explore the use of a parallel corpus of Cantonese and Mandarin Chinese (Lee, 2011) as bilingual teaching material. The corpus contains more than 8000 Cantonese-Mandarin sentence pairs; the Cantonese sentences are transcriptions of television programmes, while the Mandarin sentences are the corresponding subtitles. In addition, the Cantonese and Mandarin words in each sentence pair are aligned. An example is shown in Table 1.

While computer-assisted language learning (CALL) for Mandarin has been much investigated (e.g. Shei & Hsieh, 2012; Yang & Xie, 2013), less attention has been paid to acquisition of Cantonese as a second language. Most previous studies have focused on pronunciation (Ki, 2006; Shī, 2002; Wong, 2010; among others), while research in vocabulary acquisition has been limited to contrastive studies of the correspondence between these two languages (e.g. Zeng, 1991). This paper is the first to evaluate the use of parallel corpus for teaching Cantonese to Mandarin speakers. In a classroom experiment, we show that our corpus-based pedagogical method significantly improved the students' Cantonese proficiency.

Table 1. Examples of word-aligned Cantonese-Mandarin sentence pairs from the parallel corpus used in our study (Lee, 2011)

Cantonese	俾*béi*	你*néih*	偷*tāu*	咗 *jó*
Mandarin	被*bèi*	你*nǐ*	偷*tōu*	了*le*
Gloss	PASS	2SG	'steal'	**perfect.aspect.particle**
Translation	"[It's been] stolen by you"			
Cantonese	大件事 *daaihgihnsih*	喇 *la*		
Mandarin	糟糕 *zāogāo*	了*le*		
Gloss	'terrible'	**mood.particle**		
Translation	"[That's] terrible!"			

The Mandarin *le* has different Cantonese counterparts in different contexts. As a perfect aspect particle, its Cantonese equivalent is *jó* (top sentence); as a mood particle, however, its Cantonese equivalent is *la* (bottom sentence, Table 1).

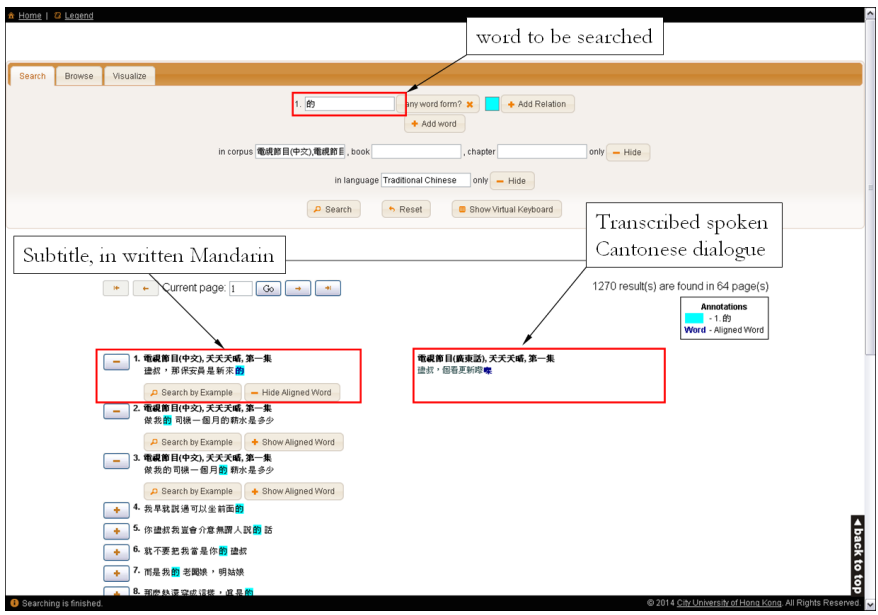
One can for example search sentences using Cantonese or Mandarin keywords, and view word alignments between Cantonese-Mandarin sentence pairs (Lee, Hui, & Yeung, 2013)

2. Experiment

2.1. Research question

Beyond the textbook, language teachers often want to employ authentic examples from contemporary media as pedagogical material in the classroom. Because Cantonese is a predominantly spoken language, it can be difficult to find such examples for Cantonese in the written form. In this study, we explore the use of a recently compiled parallel corpus of Cantonese and Mandarin Chinese (Lee, 2011) for this purpose. We have developed a web interface (Figure 1) to facilitate independent learning of Cantonese by Mandarin-speaking students. Students can retrieve sentences containing particular Mandarin or Cantonese keywords, view the Mandarin-Cantonese sentence pairs, and study the word alignments (Lee et al., 2013). In this paper, we investigate the extent to which this corpus-based method enhances the teaching of Cantonese as a second language.

Figure 1. The web interface used in the CALL session in our study



2.2. Experiment design

The evaluation took place at a 13-week course, *Cantonese Communication Skills for Putonghua Speakers*, offered at City University of Hong Kong. In total, 34 students

participated, of which 27 completed all four tasks. All were Mandarin-speaking undergraduate students. Before taking this course, most had little knowledge of Cantonese.

During the 7th week, we administered a pre-test, which contained 24 Mandarin sentences, each with one word underlined. The students were asked to translate the underlined word into Cantonese. We chose Mandarin words (e.g. the word *le* in Table 1) that had at least two different translations in Cantonese (*jó* and *la*), depending on context. Specifically, the test assessed each student on 12 Mandarin words, each appearing in two sentences requiring two different Cantonese translations. Overall, for 47.8% of these words, the students gave incorrect Cantonese translations in at least one of the two contexts; we collected these words to be used in our CALL experiment.

Two weeks later, each student completed a CALL session, using the web interface of the parallel corpus shown in Figure 1. Given a list of Mandarin words, the student was asked to search for sentences in which they appear, retrieve the original transcribed Cantonese utterance, and analyse the meanings and functions of the Cantonese words to which they were aligned. We personalised the list for each student, by randomly selecting half of the Mandarin words which the student failed to translate correctly in the pre-test (henceforth, the "CALL set"), and excluding the other half as control (the "non-CALL set").

Immediately after this session, we administered a post-test to measure the short-term effect of the session. As in the pre-test, the student was asked to translate the same 24 Mandarin words, although in different sentences and contexts. After three weeks, we administered a delayed post-test, using the same Mandarin words but again in different contexts, to measure the long-term effect.

2.3. Experimental results

A summary of our experimental results is given in Table 2. To ensure there is no significant difference in the students' previous knowledge about the words in the CALL and non-CALL sets, we first compare their performance on these two sets in the pre-test. The students correctly translated 38.0% of the words in the CALL set, and 36.8% in the non-CALL set. The difference in their performance on the two sets is not significant[3].

3. By chi-square test, $\chi^2=0.36$, $p>0.05$, $d.f.=1$

In the post-test, after the CALL session, student performance on both sets improved substantially. Even for words in the non-CALL set, which were not involved in the CALL session, the score rose to 69.4%. The students likely noticed some of these words in the sentences they browsed, and learned about their usage as a side effect. Meanwhile, the score on the CALL set increased even more sharply, to 86.7%. Hence, even with the beneficial side effect, student performance on the CALL set was significantly higher[4] than on the non-CALL set. These figures suggest that the corpus-based pedagogical approach was very effective in the short-term.

In the delayed post-test, as expected, student performance on the CALL set decreased slightly to 83.7%, while the non-CALL set increased to 74.3%. However, the score on the CALL set remained significantly better[5] than that of the non-CALL set. These results suggest that the use of the parallel corpus also improved students' Cantonese proficiency in the long-term.

Table 2. Average student score for Mandarin-to-Cantonese translation, divided into the CALL set (those words that are studied in the CALL session) and the non-CALL set (the rest)

Average score	CALL set	Non-CALL set
Pre-test	38.0%	36.8%
Post-test	86.7%	69.4%
Delayed post-test	83.7%	74.3%

Scores on the former set were significantly higher than on the latter set in the post-test and delayed post-test. In all three tests, there were 166 words in the CALL set and 144 in the non-CALL set.

3. Conclusions and future work

We have investigated the effect of a corpus-based pedagogical method for teaching Cantonese as a second language. This method centers on a personalised computer-assisted language learning session, where each student actively searched in a parallel corpus (Lee, 2011) and learned about usage of Cantonese words for which s/he previously failed to master. Compared to the pre-test, the students demonstrated significantly higher proficiency in Cantonese in the post-test; the long-term effect, as measured in the delayed post-test, is also strong.

4. By chi-square test, χ^2=0.99, p<0.0005, $d.f.$=1

5. By chi-square test, χ^2 =0.84, p<0.05, $d.f.$=1

For future work, we would like to enrich the corpus by providing pronunciation information, and to use the interface to teach Cantonese grammar.

References

Ki, W. W. (2006). *Computer-assisted perceptual learning of Cantonese tones*. Paper presented at the 14th International Conference on Computers in Education, Peking, Nov 30-Dec 4.

Lee, J. (2011). Toward a parallel corpus of spoken Cantonese and written Chinese. In *Proceedings of the 5th International Joint Conference on Natural Language Processing* (pp. 1462-1466), Chiang Mai, Thailand, November 8−13, 2011.

Lee, J., Hui, Y. C., & Yeung, C. Y. (2013). Toward a digital library with search and visualization tools. In *Proceedings of the 6th Language and Technology Conference*.

Matthews, S., & Yip, V. (2011). *Cantonese: A comprehensive grammar* (2nd ed.). London: Routledge.

Shei, C., & Hsieh, H.-P. (2012). Linkit: A CALL system for learning Chinese characters, words, and phrases. *Computer Assisted Language Learning*, 25(4), 319-338. doi:10.1080/09588221.2011.589390

Shī, Z. (2002). *Guǎngzhōu yīn Běijīng yīn duìyìng shǒucè* 廣州音北京音對應手冊 [A handbook on the correspondence between Cantonese pronunciation and Pekinese pronunciation]. Canton: Jinan University Press.

Wong, T.-S. (2010). *A pilot study on the outcome of teaching phonological correspondence in Cantonese class for Mandarin speakers*. Paper presented at the 2010 Annual Research Forum of the Linguistic Society of Hong Kong (LSHK-ARF 2010), Hong Kong. Retrieved from http://www.lshk.org/arf2010/doc/LSHK-ARF_2010_abstracts_2.0.pdf

Yang, C., & Xie, Y. (2013). Learning Chinese idioms through iPads. *Language, Learning & Technology, 17*(2), 12-23.

Zeng, Z. (1991). *Colloquial Cantonese and Putonghua equivalents* (3rd ed.) (S. K. Lai, Trans.). Hong Kong: Joint Publishing (Hong Kong) Company Limited.